Supplementary Information

Participants

Informed consent was obtained from 47 participants and their parents in 2004 (Time 1). Thirty-five participants returned for testing in 2007/8 (Time 2) and informed consent was repeated. Thirty-four completed all tests but, following testing, one participant was excluded because FSIQ was less than 70 (the threshold for the assessment of mental retardation).

The sample was selected to provide a wide range of abilities. The majority had previous educational assessments for either high or low ability. One group (18 participants) had been assessed by a trained educational psychologist because of unexpectedly poor educational achievement, particularly in reading and spelling, consistent with a diagnosis of developmental dyslexia, although their difficulties were not limited to reading. The second group (11 participants) had been reported as having high educational achievement on the basis of an entrance examination for a selective school. Four participants had not been assessed for either high ability or dyslexia. Critically, however, there is an important difference between “selecting on the basis of test taken” and “selecting on the basis of ability”. The latter was only possible once we had tested all participants on the same tests (as opposed to the different tests used in the high ability assessment and the dyslexia assessment). The results of the common assessment (the IQ test) in all our participants revealed a distribution of scores that did not differ significantly from a normal distribution, suggesting a continuum that should be representative of the general population.

In post hoc tests, we split our normally distributed sample into three ability groups: average, low and high (see Supplementary Table 1). The average ability group (n=7) were those whose FSIQ fell in the average range (80 to 119) at Time 1 and whose reading and spelling (age adjusted scores on the Wechsler Objective Reading Dimensions) at Time 1 were within the range observed for those who had not been assessed for dyslexia. The remaining participants were split into high (n=11) and low (n=14) ability groups with one subject not fitting neatly into either group. The high ability group had FSIQ of at least 120 and no dyslexia. The low ability group had FSIQ below the range of those in the high ability group and also had reading or spelling scores below the range observed for those who had not had a dyslexia assessment. The participant who did not fit into any of the groups had exceptionally high FSIQ (=135) despite poor spelling ability.

All three groups contained a wide range of changes in scores between testing points, with some individuals increasing their score and others showing either no change or a fall in score, as previously discussed in dyslexics[26-27]. Overall, the groups did not differ significantly in terms of their change in score on any of the three measures used here (although there was a trend towards a difference on VIQ, where the average group showed a larger improvement than the other two groups), nor in terms of their change in grey matter density in the areas identified in the main analysis. Post hoc correlations that focused on grey matter density in the regions of interest from the main analysis across all participants (see main text) showed that the core results of our study were significant ($p<0.05$ in 2-tailed tests) for all three ability groups, with the exception of the PIQ finding for the high ability group ($p = 0.12$, 2 tailed; $p=0.06$ one-tailed). The resulting plots are shown in Supplementary Figure 1. The consistency of the correlations in each ability groups suggests that our results are robust and should therefore generalise to other teenagers.

Behavioural tests

For cross-sectional consistency, it was necessary to use the same tests for all participants within a time point. This ensured that the inter-subject variability in the change in standardized score was not confounded with inter-subject variability in the test materials, and also ensured that any practice effects in performing the tests would be minimised (although these would have been small in any case given the 3.5 year interval between testing points). There were 9 sub-tests that were used in both the WISC and the WAIS (Table 2). These have the same form but use different stimuli to avoid ceiling and floor effects at different developmental stages. In addition, the WAIS testing included Digit Span and Matrix Design and WISC testing included Object Assembly. The scores on each sub-test are standardized according to age-specific norms to produce scaled scores. Sub-tests are allocated to either Verbal or Performance categories and the appropriate scaled scores are summed to produce totals for each category; the overall IQ measure is derived from the sum of the Verbal and Performance categories. These totals are standardized separately for VIQ, PIQ and FSIQ to produce IQ scores with a distribution of mean 100 and standard deviation 15. In all cases, the change in score was obtained by subtracting the appropriate Time 1 score from the equivalent Time 2 score. The mean time between behavioural testing and scan was 1.4 weeks (standard deviation 2.8 weeks).

Additional post hoc tests

The gap between scans was not significantly correlated with either the change in performance between tests or the change in grey matter density between tests. Males and females did not differ significantly in terms of (i) measured IQ at either test, (ii) the change in performance between tests or (iii) the change in grey matter density between tests. The correlations between Time 1 and Time 2 performance were not significantly different for males and females on any of the IQ measures.

We found no significant within year cross-sectional effects of VIQ, PIQ or FIQ on brain structure, probably due to large between subjects variation in brain structure.

Functional data

The same 33 subjects also participated in a functional study on both occasions. Full details of the paradigm, data acquisition and analysis have been reported elsewhere[23], but briefly there were eight different conditions. Four involved articulation (picture naming, reading aloud or saying “1, 2, 3” to unfamiliar Greek letters or pictures of non-objects). The other four involved a finger press response (semantic decisions on pictures of objects, semantic decisions on written words, perceptual decisions on unfamiliar Greek letters or pictures of non-objects). The paradigm allows us to segregate activity related to visual, perceptual and semantic processes, lexical retrieval, decision making, articulation and finger press responses[24]. The second level functional imaging analysis used a standard factorial analysis and identified the positive and negative effect of speech relative to right hand finger press responses. Greater activation for speech was identified in bilateral sensorimotor cortices and bilateral auditory cortices while greater activation for right hand finger presses was identified in the left sensorimotor hand area, the contralateral right superior cerebellum and a region in the midline anterior cerebellum (Lobule IV). Parts of these systems corresponded with the areas identified in the structural imaging analysis. The left motor speech area that we associated with VIQ were activated by all articulation conditions relative to finger press conditions. The midline anterior cerebellar region associated with PIQ were activated by all finger press tasks relative to articulation conditions. Figure 2 in the main text illustrates the similarity between the functional and structural neuroimaging findings. The identification of these areas in the functional analysis confirms the relevance of the areas identified in the structural analysis to the IQ measures being used. However there were no correlations between the change in activations between tests in the functional analysis and the change in grey matter density between tests in the structural analysis. Thus the structural data dissociated the effect of VIQ and PIQ on brain structure and the fMRI activation paradigm was used to assess the underlying sensorimotor functions.

Additional References

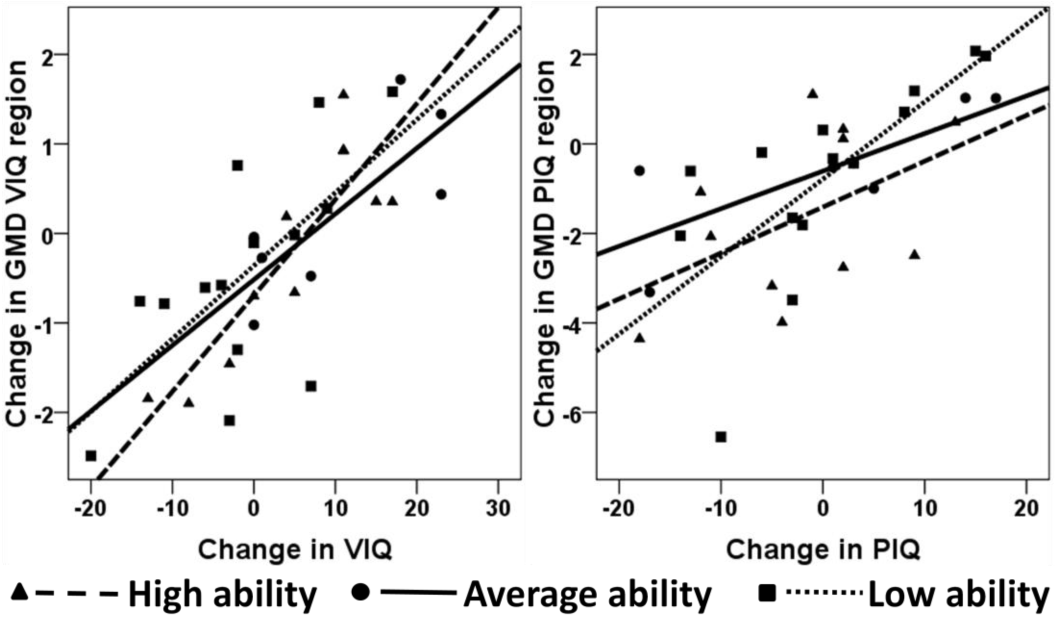
26. Siegel, L. S. & Himel, N. Socioeconimic status, age and the classification of dyslexics and poor readers: the dangers of using IQ scores in the definition of reading disability. *Dyslexia*, **4**, 90-104 (1998).

27. Stanovich, K. E. Explaining the differences between the dyslexic and the garden-variety poor reader: the phonological-core variable –difference model. *J Learn Disabil,* **21(10)**, 590-604 (1988).

Supplementary Table 1: Comparison of characteristics of high, average and low ability groups

		Age	FSIQ	VIQ	PIQ
High ability (n=11)					
Time 1	Mean (SD)	13.8 (0.9)	126 (3.8)	128 (8.8)	115 (7.4)
	Min/max	13.1/16.5	120/132	113/139	101/125
Time 2	Mean (SD)	17.4 (1.0)	126 (9.1)	133 (10.8)	113 (6.4)
	Min/max	16.6/20.2	110/143	117/150	98/124
Change	Mean (SD)	3.6 (0.1)	+0.4 (8.2)	+4.5 (9.7)	−2.1 (9.2)
(Time 2 – Time 1)	Min/max	3.5/3.7	−16/+11	−13/+17	−18/+13
Average ability (n=7)					
Time 1	Mean (SD)	14.4 (1.3)	108 (7.8)	111 (9.3)	103 (9.2)
	Min/max	13.0/16.3	94/118	100/121	90/119
Time 2	Mean (SD)	17.9 (1.3)	114 (10.8)	121 (14.8)	104 (11.2)
	Min/max	16.4/19.8	98/128	100/138	85/117
Change	Mean (SD)	3.5 (0.1)	+6.1 (9.7)	+10.3 (10.7)	+0.7 (13.7)
(Time 2 – Time 1)	Min/max	3.3/3.6	−9/+21	0/+23	−18/+17
Low ability (n=14)					
Time 1	Mean (SD)	14.1 (0.9)	101 (10.3)	101 (9.5)	102 (11.7)
	Min/max	12.6/16.0	77/114	84/115	74/116
Time 2	Mean (SD)	17.7 (0.9)	101 (6.8)	99 (7.8)	102 (7.4)
	Min/max	16.0/19.5	87/111	90/113	83/111
Change	Mean (SD)	3.6(0.2)	−0.5 (9.0)	−1.1 (9.9)	+0.1 (9.4)
(Time 2 – Time 1)	Min/max	3.3/3.9	−18/+13	−20/+17	−14/+16

SD standard deviation
See Supplementary Material for definition of groups



Supplementary Figure 1: Effect of interest in high, average and low ability groups

Correlation between change in % grey matter density and change in VIQ score (left) and PIQ (right) in the regions identified with VIQ (left) and PIQ (right) in the main analysis but shown separately for three ability groups: high (n = 11), average (n = 7) and low (n = 14) (see Supplementary Materials for details of the group definitions). Correlation coefficients between change in VIQ scores and change in grey matter density were 0.876 ($p<0.01$) for high ability, 0.797 ($p<0.05$) for average ability and 0.660 ($p<0.05$) for low ability groups respectively. For PIQ, the corresponding effects were 0.492 (not significant) for high ability, 0.788 ($p<0.05$) for average ability and 0.715 ($p<0.01$) for low ability groups.